Open comments on the Task Force SIRS report: Scholarly Infrastructures for Research Software (EOSC Executive Board, EOSCArchitecture)

Teresa Gomez-Diaz (CNRS/LIGM), Tomas Recio (University of Cantabria)

Contact: Teresa.Gomez-Diaz@u-pem.fr, Tomas.Recio@unican.es

November 2nd 2020, V1

1 Foreword

The goal of this document is to openly contribute with our comments to the EOSCArchitecture report: *Scholarly Infrastructures for Research Software (SIRS)*, draft version dated October 2020. This SIRS draft report is open for consultation as a Google document [1], as has been announced at the EOSC Symposium (19-22 October 2020, Online) [2].

Political evolutions in digital policy have been recently announced by Ursula von der Leyen, President of the European Commission. They are designed to enhance Europe's strategic autonomy [10]. In this context, and in order to contribute to and to support the Europe's Digital sovereignty, we have avoided the use of Google accounts, and therefore the direct editing of the above mentioned Google document. Our participation to this EOSC effort takes then the form of the present document, that will be available in Zenodo.

2 Comments

In this section we propose a list of comments to the SIRS draft report. Each comment is associated to one section or subsection of the draft.

Please note that some short extracts of the SIRS report have been included here and are, thus, out of context. We recommend the consultation of the original text, as maybe some mistakes or misinterpretations could have been unintentionally introduced in the present text.

Section 2.1 Scope and goals - Research software definition This section 2.1 introduces the concept of research software used in the report as follows:

> *the term "research software" may carry very different meanings in different research communities: in this report, we will use this term simply to designate software that researchers in any discipline may feel the need to have scholarly infrastructure support for, no matter if it is considered a tool, a result or an object of study.*

Please note that this definition does not imply any difference between the concepts of software and research software, as this research software presentation could easily include any version of the Windows operating system, Matlab and many other commercial software. For example a history researcher could feel that all software developed since 1960 should be preserved for further studies.

[1] https://docs.google.com/document/d/1yObRCR7COQctpjMdg-rNenRZ7ZeUZ-u0iyvdpPRK598

[2] https://www.eoscsecretariat.eu/eosc-symposium-2020, recording available at https://www.youtube.com/watch?v=U_sxfV0kjEg

It seems to us that to define EOSC infrastructures and services based in this view of research software is a task that requires some essential precisions. Otherwise, using strictly the above research software definition, how could EOSC teams design adapted services? For which target community? In particular, it is necessary to pay careful attention when dealing with software produced by private companies. It seems to us that these questions are not correctly presented or are missing in the report.

On the other hand, EOSC is presented in [1] as follows:

> *The EOSC will be a fundamental enabler of Open Science and of the digital transformation of science, offering every European researcher the possibility to access and reuse all publically funded research data in Europe, across disciplines and borders.*

In addition, we have proposed the following definition for Open Science in this recent preprint [6]:

> *Open Science is defined as the political and legal framework where research outputs are shared and disseminated in order to be rendered visible, accessible, reusable.*

According with this Open Science vision, we propose another research software definition as follows:

> *research software is a well identified set of code that has been written by a (again, well identified) research team. It is software that has been built and used to produce a result published or disseminated in some article or scientific contribution. Each research software encloses a set (of files) that contains the source code and the compiled code. It can also include other elements as the documentation, specifications, use cases, a test suite, examples of input data and corresponding output data, and even preparatory material.*

You can find this definition and all the considerations for its proposition in section 2.1 of [5].

Thus, the research software definition in [5] places correctly research software as a research output, usually produced within publicly funded research, and the role of EOSC efforts correspond to the definition and implementation of infrastructures and services to render research outputs, including research software, *visible, accessible, reusable.*

One of the authors of the present document provides with a good example to further study this research software concept. T. Recio is currently studying automatic proving of geometric theorems through dynamic geometry software, and comparing existing work done with the computer language Logo [3] and with the Cabri-geometry (commercial) mathematical application [4]. The outputs of this research are currently being implemented in Geogebra [5]. Under the SIRS draft report definition, all these three objects should be considered as research software, but with the definition on [5], only the last one fits well in the concept. Research software is then the result of the ongoing scientific work, and not the historical tool or the commercial software which are under study. The output and the producer are therefore correctly identified.

The SIRS TF should clarify if historical tools and commercial software should be considered as part of the objects for which EOSC should provide infrastructure and services, and how commercial software should be dealt with.

Section 2.2 Infrastructures participating in the Task Force (TF) This section presents summary sheets introducing the nine infrastructures that are represented in the SIRS TF: three for the Archives category (HAL, Software Heritage, and Zenodo), three for the publisher category (Dagstuhl, eLife, and IPOL), and three in the aggregators category (OpenAIRE, ScanR, and swMath).

We would like to suggest, for all these nine infrastructures, to add the following information, that could be presented in an homogeneous way:

— Funders, their role: do they provide physical hosting facilities, numerical resources, human resources, other funding... distinguish between public and private funding
— Governance: with a link to the list of persons involved in the governance and their organization
— Teams: with a link to the list of the teams involved in the infrastructure and their organization
— Services: list of services provided
— Target public: specific research communities, entities...

[3] https://en.wikipedia.org/wiki/Logo_(programming_language)
[4] https://swmath.org/software/4928
[5] https://swmath.org/software/4203

- Legal framework: links to the texts providing the conditions in which the services are provided, about the personal data processing...
- Software: link to the software that is employed to run the platforms and services
- Data storage: which are the entities involved in the data storage, and where are the servers that store the whole data

This Section 2.2 also indicates that:

> *In the context of this report we use the term [...] 'Publishers' are organizations that prepare submitted research texts, possibly with associated source code and data, to produce a publication and manage the dissemination, promotion, and archival process. Software and data can be part of the main publication, or assets given as supplementary materials depending on the policy of the journal. In addition, publishers implement a process for ensuring the quality of the accepted research material (usually peer Review), which is carried out by a subject-specific community of experts.*

The report could provide further clarification about if these publications include Data papers and Software papers, and if the software mentioned in this paragraph corresponds to the research software defined in the section 2.1 of the SIRS report (see the above comment on this section). For those that are unfamiliar with Data or Software papers, examples of scientific journals publishing biodiversity-related data papers can be found in the Global Biodiversity Information Facility (GBIF) web site[6]. Examples of scientific journals publishing Software papers are listed in the Software Sustainability Institute (SSI) web site[7] and some have been analyzed in the section 2.4 entitled *Publication of research software* of [5].

Section 3.1 Survey on Related Initiatives and Related Works The SIRS report indicates that:

> *it seems that general awareness about the importance of software as a research output has started growing only very recently, around 2010, in particular as a byproduct of the reproducibility crisis (Barnes, 2010; Borgman et al., 2012; Colom et al., 2015; Konrad Hinsen, 2013; Rougier et al., 2017; Stodden et al., 2012).*

Please note that, at least in France, there have been older initiatives. The Project PLUME (2006-2013), launched by the UREC CNRS unit, has studied research software and its dissemination conditions. It has also published research software descriptions and *validated software descriptions* [2, 4, 5]. PLUME has also provided training and support services at national level, see the section *Patrimoine logiciel d'un laboratoire*[8].

At the time, the term research software was not really existing or maybe its use was not widely extended, so the terms used by the PLUME project where *logiciel d'un laboratoire*, that is, software produced in a French research lab, and the term *dév. Ens Sup - Recherche* or *dév. ESR*, the short forms of *développements de l'enseignement supérieur et la recherche*, that is, developments realised in the Higher Education and Research community.

Section 3.1.3 Aggregators The SIRS report indicates that:

> *Another remarkable example is the catalog built by the Plume project in order to collect information about software that is useful for research activities (Plume, 2013): it maintains a collection of over 400 entries manually curated about software projects that are successfully deployed and in use in at least three different research laboratories.*

Please note that the PLUME Project has published 406 *validated software descriptions*[9], 358 research software descriptions in French[10] and 116 research software descriptions in English[11] between 2007 and 2013. This catalogue has been produced with publication and peer review procedures (not with automatic

[6] `https://www.gbif.org/data-papers`

[7] `https://www.software.ac.uk/resources/guides/which-journals-should-i-publish-my-software`

[8] `https://projet-plume.org/patrimoine-logiciel-laboratoire`. Note that *patrimoine logiciel* translates to the English term *software heritage.*

[9] `https://projet-plume.org/types-de-fiches#logiciel_valide`

[10] `https://projet-plume.org/fiches_dev_ESR`

[11] `https://projet-plume.org/en`

aggregation procedures), which have involved 2220 contributors, including writers, reviewers and PLUME team members. The project is frozen and the catalogue is not maintained any more [4].

Section 3.2.2 Publishers The SIRS report mentions that

> *Over the past few years several publishers have led the effort in the transition towards open access as the predominant model of publication for scholarly outputs. This also paves a path for fair and affordable conditions from the start for the dissemination of software, but support for software outside of specialist journals is still limited.*

As presented in [7], the identified four key functions needed by scholarly publishing are:

registration *to establish that work had been undertaken by individuals or groups of researchers at a particular time, and thus their claim to precedence;*

certification *to establish the validity of the findings;*

dissemination *to make scholarly works and their findings accessible and visible;*

preservation *to ensure that the 'records of science' are preserved, and remain accessible, for the long term.*

The SIRS report could clarify what means software support provided by publishers.

Please also note that the section 2.4 of [5] entitled *Publication of research software* does provide another partial panorama of the research software publication world (as defined in this publication). Besides, the section 2.5 *Referencing and citation of research software* provides further studies on the reference and the citation issues for research software.

Section 3.2.3 Aggregators The SIRS report mentions here the OpenAIRE Research Graph. Please note that there is also the Software Heritage graph Dataset presented in [11] and funded by the Horizon 2020 research and innovation programme under grant agreement No 825328 (FASTEN). The Software Heritage graph dataset is available in multiple formats, which includes a public instance on Amazon Athena interactive query service for ready-to-use powerful analytical processing.

The SIRS report could provide further insight over comparisons between these two different graphs and the possible interactions between these two works.

Section 3.3 Best Practices and Open Problems This section includes three tables which are surely the result of extended discussions inside the working groups. Please note that for readers that have not been involved in the debates, these tables (and the terms and expressions used inside) need further explanation. For example the table in section 3.3.1 Best Practice Principles for Archives includes a last column entitled Priorities indicating levels of Development, Adoption, Research, Harmonization. These terms remain unclear and they could be put in a more precise context.

Section 3.3.1 Best Practice Principles for Archives This section indicates that

> *one does not need to reinvent the wheel, the archival community should agree on an overall architecture to integrate existing infrastructures.*

We would like to express our agreement with this idea, which we have found very inspiring. We are not experts in archiving services nor data treatment or management, but it seems that archiving software source code may find many common issues with data archiving, and thus, both services could be compared and put into perspective. The SIRS TF team could consult with members of the EOSC-HUB project [12] like BSC, CSC, CINES (and maybe others) and exchange about common gaps and best practices, in order to not to reinvent the wheel.

[12] `https://www.eosc-hub.eu/partners`

Section 3.4.2 Identifiers This section presents the need of proper identification for software artifacts and presents the SoftWare Heritage persistent identifiers (SWHIDs) as a candidate for solution.

Please note that there are other EOSC teams working on persistent identifiers (PIDs) issues [8, 9] as well as FREYA [13], a whole EC funded project.

The work proposed in the SIRS report as a solution could be more connected with these others EC funded ongoing efforts, and collaborative work should be carried on in order to propose and adopt consensual solutions.

Section 3.4.3 Quality and Curation The table mentions the *Evaluation of source code.*

Please note that the CDUR procedure to evaluate research software has been proposed in [5]. CDUR comprises four steps introduced as follows: Citation, to deal with correct RS identification, Dissemination, to measure good dissemination practices, Use, devoted to the evaluation of usability aspects, and Research, to assess the impact of the scientific work.

The SIRS report could consider to include the CDUR procedure in the list of methods to assess research software quality, as CDUR includes the evaluation of the research software source code in the Use step.

Sections 3.4.4 Metrics, 3.4.5 Guidelines, 3.4.6 Tools and Workflows The SIRS report could provide further details for proposed guidelines, metrics and tools and workflows.

Section 4.1.1. Archive This section mentions

> *1. Layer 0: universal archive specifically designed for software source code*
> *— proactive archival of all software source code (including all dependencies of research software)*
> *[...] 2. Layer 1: scholarly repositories*
> *— explicit deposit by identified individuals...*

Here we find the problem of the definition of research software (see comment on section 2.1), and the importance of having well defined objects in order to design sound services and infrastructures to deal with them.

The Layer 0 considers the archiving of every existing software and in Layer 1 we find a more “usual” research software object, as the individuals do the deposit of their own production in the scholarly repositories.

Section 5.1.1 Interactions The SIRS report mentions that

> *it is important to ensure a vertical interconnection between an universal software archive and scholarly repositories, for the latter to feed the universal archive (see Figure 5). This requires engineering and funding for the development of proper adaptors.*

The SIRS report could provide further insight on the goals of the Scholarly Infrastructures for Research Software studied in the report concerning their contribution to feed universal archives.

[13] `https://www.project-freya.eu/en/about/mission`

3 Final Remarks

Finally we conclude this short report with some further questions to be considered by EOSC decision makers, as we think that some of the issues raised by the SIRS draft need extended and in-depth reflection.

Definition The concept of research software is essential for a sound design of infrastructures and services that will deal with this research output.

Software Management Plan It is now widely accepted that a Data Management Plan (DMP) is an important tool when dealing with research data, and DMPs are usually required by funders. Tools for Software Management Plans are also available, see for example [3].

Services The SIRS report considers three kind of existing infrastructures: Archives, Publishers, Aggregators. The issues studied for software and/or for research software are *Archive, Reference, Describe, and Credit*, as mentioned in section 2.1.1. On the other hand, if EOSC's goals are to render research software *visible, accessible, reusable* there is also need for services like, for example, *search, testing, and retrieval interfaces.* A *search engine* is mentioned in the SIRS report as a long term perspective (see the Executive Summary and the section 5.3.1 Advanced Technology Development). But, as far as we understand, it is the first service that researchers will ask for.

User-centric EOSC As well as the services that will be provided, there is the question of the interactions with foreseen users. Relevant members of the EOSC construction have signed a joint statement [14] to signal *the importance of making EOSC relevant for scientific communities and researchers.*

Architecture EOSC is already a complex system with several key actors of distinct nature. Interactions and collaborations among all the different components should be designed to facilitate the user approach to EOSC.

Co-Funding As mentioned in [10]:

> *Reliable digital infrastructure and services are critical in today's society, as the coronavirus crisis has highlighted. A range of initiatives have been proposed or are already under discussion at EU level to accelerate the digitalisation process and enhance Europe's strategic autonomy in the digital field.*

In this context, EOSC decision makers should consider the co-funding of infrastructures already funded by non-EU tech companies in a transparent way.

Acknowledgements We would like to thank the SIRS TF team for such an inspiring work.

[14] `https://www.openaire.eu/eosc-a-tool-for-enabling-open-science-in-europe`. See the statement on the EOSC Secretariat website at `https://www.eoscsecretariat.eu/eosc-liaison-platform/post/research-oriented-services-trust-collaboration-sustainability-key`.

References

[1] European Commission. Commission staff working document. Implementation Roadmap for the European Open Science Cloud. Brussels, 14.3.2018 SWD(2018) 83 final. https://ec.europa.eu/transparency/regdoc/rep/10102/2018/EN/SWD-2018-83-F1-EN-MAIN-PART-1.PDF

[2] T. Gomez-Diaz. Article vs. Logiciel : questions juridiques et de politique scientifique dans la production de logiciels. 1024 - Bulletin de la société informatique de France, N. 5 - March 2015, http://www.societe-informatique-de-france.fr/wp-content/uploads/2015/04/1024-5-gomez-diaz.pdf. First version initially published in the platform of the PLUME project, October 2011, https://projet-plume.org/ressource/article-vs-logiciel

[3] T. Gomez-Diaz, G. Romier. Research Software management Plan Template V3.2. Projet PRESOFT, Bilingual document (FR/EN), 2018, https://hal.archives-ouvertes.fr/hal-01802565. Also available on the CNRS DMP OPIDoR service, https://opidor.fr/plan-de-gestion-de-logiciel-de-la-recherche-projet-presoft-disponible-sur-dmp-opidor/

[4] T. Gomez-Diaz. Le Projet PLUME et le paysage actuel des logiciels de la recherche dans la science ouverte. Zenodo, Preprint, 2019. https://zenodo.org/record/2591474

[5] T. Gomez-Diaz and T. Recio, On the evaluation of research software: the CDUR procedure [version 2; peer review: 2 approved]. F1000Research 2019, 8:1353, https://doi.org/10.12688/f1000research.19994.2

[6] T. Gomez-Diaz and T. Recio, A policy and legal Open Science framework: a proposal. Zenodo, Preprint, 2020, https://zenodo.org/record/4075106

[7] Jean-Claude Guédon et al. Future of Scholarly Publishing and Scholarly Communication, Report of the Expert Group to the European Commission, January 2019, https://www.eosc-portal.eu/sites/default/files/KI0518070ENN.en_.pdf

[8] Second draft Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC), V2 May 2020, https://zenodo.org/record/3780423

[9] PID Architecture for the EOSC, V0.3, 2020-10-15, Draft for consultation, https://docs.google.com/document/d/1T-bpNsmuxQewsLq48XTyUJoe0lsV7poaXohpgDo9W34

[10] Tambiama Madiega, European Parliamentary Research Service (EPRS). Digital sovereignty for Europe. EPRS Ideas Paper, PE 651.992, July 2020, https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/651992/EPRS_BRI(2020)651992_EN.pdf

[11] Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli. The Software Heritage Graph Dataset: Public software development under one roof. Proceedings of the 16th International Conference on Mining Software Repositories, pp. 138-142, IEEE Press, 2019. https://www.softwareheritage.org/wp-content/uploads/2020/01/msr-2019-swh.pdf